

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平11-154058

(43)公開日 平成11年(1999)6月8日

(51)Int.Cl.⁶

G 0 6 F 3/06

識別記号

3 0 5

5 4 0

F I

G 0 6 F 3/06

3 0 5 C

5 4 0

審査請求 未請求 請求項の数 6 O L (全 9 頁)

(21)出願番号 特願平9-319359

(22)出願日 平成9年(1997)11月20日

(71)出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72)発明者 中村 寛男

東京都府中市東芝町1番地 株式会社東芝

府中工場内

(72)発明者 笹本 享一

東京都府中市東芝町1番地 株式会社東芝

府中工場内

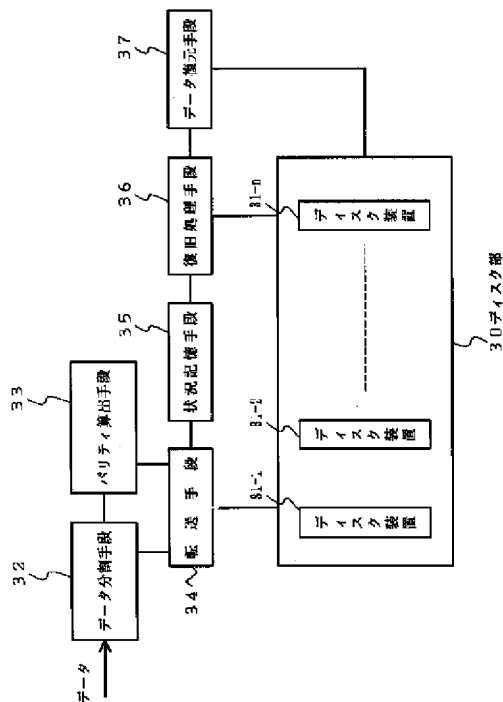
(74)代理人 弁理士 本田 崇

(54)【発明の名称】 ディスクアレイ装置及びデータ保守方法

(57)【要約】

【課題】 データ書込途中に電源断が生じて、データを適切に復元する。

【解決手段】 データを記憶する複数のディスク装置31-1～31-nと、記憶すべきデータを所定の長さに分割するデータ分割手段32と、分割した所定数のデータにより構成されるパリティグループについてパリティデータを算出するパリティ算出手段33と、前記分割データ及びパリティデータを前記複数のディスク装置31-1～31-nへ転送する転送手段34と、前記記憶すべきデータの格納先アドレス、サイズ、転送状況に関する情報を検出し記憶する状況記憶手段35と、復旧の際に前記状況記憶手段35の情報を検索して転送が未完を検出し、当該格納先アドレス及びサイズに基づき転送が未完となっている領域に係るパリティグループについて再度パリティデータを作成して、該当するディスク装置に書き込む復旧処理手段36とを具備する。



【特許請求の範囲】

【請求項1】 データを記憶する複数のディスク装置と、
記憶すべきデータを所定の長さに分割するデータ分割手段と、
分割した所定数のデータにより構成されるパリティグループについてパリティデータを算出するパリティ算出手段と、
前記分割されたデータ及びパリティデータを前記複数の所要のディスク装置へ転送する転送手段と、
前記記憶すべきデータの格納先アドレス、サイズ、転送状況に関する情報を検出して記憶する状況記憶手段と、
復旧の際に前記状況記憶手段に記憶された情報を検索して転送が未完となっている場合には、当該格納先アドレス及びサイズに基づき転送が未完となっている領域を検出し、この領域のパリティグループについてディスク装置の記憶内容に基づきパリティデータを作成して、該当するディスク装置に書き込む復旧処理手段と、
を具備することを特徴とするディスクアレイ装置。

【請求項2】 複数のディスク装置のいずれかに障害が発生した場合には、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元するデータ復元手段を具備することを特徴とする請求項1に記載のディスクアレイ装置。

【請求項3】 複数のディスク装置の特定の1台が、パリティデータを記憶するためのパリティディスクとなっていることを特徴とする請求項1または請求項2に記載のディスクアレイ装置。

【請求項4】 記憶すべきデータが与えられると分割し、
上記分割したデータに関するパリティグループについてパリティデータを算出し、
上記分割したデータ及びパリティデータを複数のディスク装置の所要のディスク装置に転送し、
前記記憶すべきデータの格納先アドレス、サイズ、転送状況に関する情報を検出して記憶し、
復旧の際に前記記憶された情報を検索して転送が未完となっている場合には、当該格納先アドレス及びサイズに基づき転送が未完となっている領域を検出し、この領域のパリティグループについてディスク装置の記憶内容に基づきパリティデータを作成して、該当するディスク装置に書き込むことを特徴とするデータ保守方法。

【請求項5】 複数のディスク装置のいずれかにおいて障害が発生したか否かを検出し、
障害が発生した場合には、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元することを特徴とする請求項4に記載のデータ保守方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、複数のディスク装置により構成されるディスクアレイ装置及びデータ保守方法に関するものである。

【0002】

【従来の技術】従来、RAID (Redundancy Arrays of Inexpensive Disks) 法によって構成されたディスクアレイ装置が知られている。図8には、RAIDレベル3によるディスクアレイ装置の構成例が示されている。ディスクアレイ装置2には、SCSI (Small Computer System Interface) 等のインタフェースバス12を介してホスト計算機(コンピュータ)1とのデータ送受を行うためのインタフェース(IF)7が備えられている。インタフェース7からはバス14が延びており、バス14には、装置全体を制御する制御部20、リード・ライトに係るデータを保持するためのデータバッファ10、パリティ計算を行うパリティ回路9、複数の(ここでは、5台の)ディスク装置(ここでは、HDD)との接続を行うインタフェース8-0~8-4が備えられている。

【0003】インタフェース8-0~8-4には、SCSI等のインタフェースバス13-0~13-4を介してHDD(ハードディスクドライブ)11-0~11-4が接続されている。このHDD11-0~11-3には、データが所定のサイズに分割されて転送されて記憶され、HDD11-4には、上記HDD11-0~11-3の対応領域のデータの排他的論理和(EX-OR)をとったパリティデータが記憶されるように構成されている。

【0004】図8において、データD0~D3がそれぞれHDD11-0~11-3の対応する領域に格納されているものとする、パリティデータPは、 $P = D0 (EX-OR) D1 (EX-OR) D2 (EX-OR) D3$ にて得られる。そして、例えば、HDD11-2に障害が発生したときには、HDD11-2に記憶されていたデータD2は、 $D2 = D0 (EX-OR) D1 (EX-OR) D3 (EX-OR) P$ にて復元されて得られる。

【0005】別の例を説明すると、図9に示されるように、HDD11-0~11-3の対応する領域N、N+1、N+2、N+3が512バイトの分割サイズを持ち、それぞれに、データ55…、AA…、99…、CC…(合計2048バイト)が記憶されるときには、パリティPは、 $55… (EX-OR) AA… (EX-OR) 99… (EX-OR) CC… = AA…$ と得られ、これがHDD11-4に書き込まれる。

【0006】そして、上記領域N、N+1、N+2に対し1536バイトのデータが、図10に示されるように、それぞれ、77…、EE…、33…と512バイトに分割されて書き込まれたときには、制御部20は、領域N、N+1、N+2、N+3のデータについて再度パ

リティデータを作成し直す。つまり、新たなパリティデータPMは、 $PM = 77 \cdots (EX-OR) EE \cdots (EX-OR) 33 \cdots (EX-OR) CC \cdots = 66 \cdots$ として得られ、これがHDD11-4に書き込まれる。

【0007】上記に対し、図11に示されるように、HDD11-0~11-3の対応する領域N、N+1、N+2、N+3に対し512バイトに分割されたデータ77…、EE…、33…を書き込む際に、電源障害等が発生し、HDD11-2に対しデータ33…の書き込みができず、正常に書き込みが終了しなかったものとする。この場合には、領域N、N+1、N+2には、データ77…、EE…、33…が書き込まれるが、領域N+3には、元のデータCC…が書き込まれた状態のまま残っており、また、HDD11-4の対応領域には、元のパリティデータがAA…が記憶されたままとなっている。

【0008】

【発明が解決しようとする課題】このような図11の状態において、例えば、図12に示すようにHDD11-3に障害が発生し、係る状態において、HDD11-3の領域N+3に対する読み出し要求があると、障害が発生しているHDD11-3を除いたディスクの対応領域のデータを用いて復元が行われる。つまり、領域N、N+1、N+2のデータ77…、EE…、33…と、パリティデータAA…を用いて、復元データ $= 77 \cdots (EX-OR) EE \cdots (EX-OR) 33 \cdots (EX-OR) AA \cdots = AA \cdots$ が復元され、HDD11-3に実際に格納されていたデータCC…と異なるデータAA…が復元されることになる。

【0009】本発明は上記のようにデータの書き込み途中に生じる電源断等の異常が生じた場合に、データを適切に復元することができなくなるという問題点に鑑みなされたもので、その目的は、データの書き込み途中に生じる電源断等の異常が生じた場合にも、データを適切に復元することができるディスクアレイ装置及びデータ保守方法を提供することである。

【0010】

【課題を解決するための手段】本発明の請求項1に係るディスクアレイ装置は、データを記憶する複数のディスク装置と、記憶すべきデータを所定の長さに分割するデータ分割手段と、分割した所定数のデータにより構成されるパリティグループについてパリティデータを算出するパリティ算出手段と、前記分割されたデータ及びパリティデータを前記複数の所要のディスク装置へ転送する転送手段と、前記記憶すべきデータの格納先アドレス、サイズ、転送状況に関する情報を検出して記憶する状況記憶手段と、復旧の際に前記状況記憶手段に記憶された情報を検索して転送が未完となっている場合には、当該格納先アドレス及びサイズに基づき転送が未完となっている領域を検出し、この領域のパリティグループについてディスク装置の記憶内容に基づきパリティデータを作

成して、該当するディスク装置に書き込む復旧処理手段とを具備することを特徴とする。これにより、復旧の際に転送が未完となっている場合には、当該格納先アドレス及びサイズに基づき転送が未完となっている領域を検出され、この領域のパリティグループについてディスク装置の記憶内容に基づきパリティデータが作成されて、該当するディスク装置に書き込みが行われる。

【0011】本発明の請求項2に係るディスクアレイ装置では、複数のディスク装置のいずれかに障害が発生した場合には、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元するデータ復元手段を具備することを特徴とする。これにより、複数のディスク装置のいずれかに障害が発生した場合には、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元しても適切な復元がなされる。

【0012】本発明の請求項3に係るディスクアレイ装置では、複数のディスク装置の特定の1台が、パリティデータを記憶するためのパリティディスクとなっていることを特徴としており、これにより所定のディスク装置にパリティデータが記憶されることになる。

【0013】本発明の請求項4に係るデータ保守装置は、記憶すべきデータが与えられると分割し、上記分割したデータに関するパリティグループについてパリティデータを算出し、上記分割したデータ及びパリティデータを複数のディスク装置の所要のディスク装置に転送し、前記記憶すべきデータの格納先アドレス、サイズ、転送状況に関する情報を検出して記憶し、復旧の際に前記記憶された情報を検索して転送が未完となっている場合には、当該格納先アドレス及びサイズに基づき転送が未完となっている領域を検出し、この領域のパリティグループについてディスク装置の記憶内容に基づきパリティデータを作成して、該当するディスク装置に書き込むことを特徴とする。これにより、複数のディスク装置のいずれかに障害が発生した場合に、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元して適切な復元が可能となる。

【0014】本発明の請求項5に係るデータ保守方法は、複数のディスク装置のいずれかにおいて障害が発生したか否かを検出し、障害が発生した場合には、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元することの特徴とする。これにより、複数のディスク装置のいずれかにおいて障害が発生した場合にデータの復元がなされる。

【0015】

【発明の実施の形態】以下添付の図面を参照して本発明の実施の形態に係るディスクアレイ装置及びデータ保守方法を説明する各図において、同一の構成要素には同一

の符号を付して重複する説明を省略する。図1には、本発明の実施の形態に係るディスクアレイ装置2Aの構成図が示されている。この装置2Aは、データを記憶する複数のディスク装置31-1～31-nからなるディスク部30を有し、このディスク部30に対し記憶すべきデータを所定の長さに分割するデータ分割手段32と、分割した所定数のデータにより構成されるパリティグループについてパリティデータを算出するパリティ算出手段33とを備える。更には、データ分割手段32により分割されたデータ及びパリティ算出手段33により作成されたパリティデータを複数の所要のディスク装置31-1～31-nへ転送する転送手段34が設けられている。

【0016】更に、ディスク部30の所要のディスク装置31-1～31-nへ記憶すべきデータの格納先アドレス、サイズ、転送状況に関する情報を検出して記憶する状況記憶手段35と、復旧の際に上記状況記憶手段35に記憶された情報を検索して転送が未完となっている場合には、当該格納先アドレス及びサイズに基づき転送が未完となっている領域を検出し、この領域のパリティグループについてディスク装置31-1～31-nの記憶内容に基づきパリティデータを作成して、該当するディスク装置に書き込む復旧処理手段36が備えられている。また、上記複数のディスク装置31-1～31-nのいずれかに障害が発生した場合には、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元するデータ復元手段37を具備する。

【0017】上記ディスクアレイ装置2Aの実際の構成例を図2に示す。この構成は、図8に示したディスクアレイ装置2と同一の構成要素に同一の符号を付して示すように基本的に同一であるが、バス14には、マイクロプロセッサ3用のCPUバス16との間のプロトコル変換を行うプロトコル変換部5が接続され、マイクロプロセッサ3がメモリ4内のプログラムやデータを用いて、上記図1に示した各手段として機能する点が特徴となっている。また、マイクロプロセッサ3は不揮発メモリ6に、ディスク部30に対応するHDD11-0～11-4に記憶すべきデータの格納先アドレス、サイズ、転送状況に関する情報を検出して記憶し、状況記憶手段35として用いる。

【0018】図3には、上記状況記憶手段35に係るメモリテーブルの構成例が示されている。つまり、メモリテーブルには、ホスト計算機1から送られ、HDD11-0～11-4に書き込まれるデータのデータ番号（書き込み中データ1～6、・・・）、格納先アドレス（1～6、・・・）、転送サイズ（1～6、・・・）、転送の完了／未完了に係る転送状況（1～6、・・・）が記憶される。

【0019】そして、マイクロプロセッサ3は、メモリ

4に格納されている図4、図5に示されるフローチャートに対応するプログラムを用いて図1の各手段として動作するので、以下にこれを説明する。ホスト計算機1からデータの書き込み要求がなされると、マイクロプロセッサ3は、これをインタフェース7プロトコル変換部5を介して受け取り（S1）、データを受けてデータバッファ10へ格納した後に、当該データを一定のサイズ（例えば、512バイト）に分割して、該当領域のパリティグループについてパリティデータを作成する（S2）。具体的には、図10に示したように、1536バイトのデータが与えられたときには、それぞれ、77…、EE…、33…と512バイトに分割し、かつ、パリティグループであるHDD11-3のデータCC…を読み出し、パリティデータPM=77…(EX-OR)EE…(EX-OR)33…(EX-OR)CC…=66…を得る。

【0020】次に、マイクロプロセッサ3は、ホスト計算機1から要求のあったデータの書き込み要求に係るデータに関し、格納先アドレス、転送サイズ、転送状況（初期値は、「未完了」）を図3のメモリテーブルに記憶する（S3）。例えば、図10に示すように書き込みを行うときには、格納先アドレス（先頭アドレス）は「N」であり、転送サイズは「1536バイト」であり、転送状況の初期値は、「未完了」とされる。次に、マイクロプロセッサ3は、該当するHDDに対しデータ書き込み要求を送出しデータ書き込みを開始し（S4）、当該データの書き込みが完了したかを検出する（S5）。

【0021】上記ステップS5において、一連のデータ（図10の例では、1536バイト）の書き込みが完了すると、図3のメモリテーブルの該当するエリアの転送状況を「未完了」から「完了」へと変更し（S6）、次の処理（ホスト計算機1からのデータの書き込み要求に係る処理）を実行する（S7）。

【0022】上記に対し、ステップS5においてデータをHDD11-0～11-4に格納しているときに、停電等の電源断が生じたときには、動作が停止し、電源投入等による復旧がなされると、図5に示されるフローチャートのプログラムが実行される。即ち、当該図5のフローチャートのプログラムのスタート等の立ち上げ時初期処理がなされ（S11）、次いでマイクロプロセッサ3は不揮発メモリ6内のメモリテーブルを検索して転送状況が「未完了」となっているデータがあるか否かを検出する（S12）。「未完了」がないときには、当該ディスクアレイ装置2Aの運用を開始する（S13）。つまり、ホスト計算機1からの要求に応じてデータの格納または読み出しを行う。

【0023】一方、転送状況が「未完了」となっているデータがある場合には、当該領域の格納先アドレス、転送サイズより、HDD11-0～11-4において書き

込みが行われなかったパリティグループの領域を検索し、当該パリティグループのパリティデータを再作成してパリティディスクであるHDD11-4へ当該パリティデータを書き込み(S14)、ステップS12へ戻って処理を続ける。

【0024】例えば、図11に示されるように、HDD11-0~11-3の対応する領域N、N+1、N+2、N+3に対し512バイトに分割されたデータ77…、EE…、33…を書き込む際に、電源障害等が発生し、HDD11-2に対しデータ33…の書き込みができず、正常に書き込みが終了しなかったものとする。このときには、この一連のデータのデータ番号が「1」であるとき、マイクロプロセッサ3によって不揮発メモリ6のメモリテーブルには、図6(a)に示されるように、格納先アドレスが「N」、転送サイズが「1536」、転送状況が「未完了」とされる。

【0025】上記のようなメモリテーブルの内容を図5のフローチャートにおけるステップS12にて検出すると、上記格納先アドレス「N」に対応してHDD11-0の領域Nに関するパリティグループのデータ77…、EE…、99…、CC…を読み出し、これらに基づきパリティデータPA=77…(EX-OR)EE…(EX-OR)99…(EX-OR)CC…を演算し、パリティデータPA=CC…を得て、これをHDD11-4の該当領域に記憶する。この結果HDD11-0~11-4には、図6(b)に示されるようにデータが格納される。

【0026】このような図6(b)の状態において、例えば、HDD11-3に障害が発生し、係る状態において、HDD11-3の領域N+3に対する読み出し要求があると、障害が発生しているHDD11-3を除いたディスクの対応領域のデータを用いて復元が行われる。つまり、領域N、N+1、N+2のデータ77…、EE…、99…と、パリティデータCC…を用いて、復元データ=77…(EX-OR)EE…(EX-OR)99…(EX-OR)CC…=CC…が復元され、HDD11-3に実際に格納されていたデータCC…と同一のデータCC…が復元されることになる。

【0027】斯して、HDD11-0~11-4に対するデータ書き込みの途中において停電等の電源断が生じて、データ書き込みが未完了となったときにも、未完了のデータ書き込みに係るパリティグループに関しパリティデータが適切なものに変更される。このため、その後にHDD11-0~11-4のいずれかに障害が生じて、障害に係るHDDのデータを適切に復元することができることになる。なお、上記の実施の形態においては、端部に配置されたHDD11-4にパリティデータを書き込むことにしたが、複数のHDDの内のいずれかにパリティデータを書き込むように構成した実施の形態が存在する。従って、特定の1つのHDDをパリティデ

ィスクとしなくとも良く、未完了のデータ書き込みに係るパリティグループに関しパリティデータを再作成して、所定の手法により決まるHDDにこのパリティデータを書き込むように構成することのできる。係る場合にも、障害に係るHDDのデータを適切に復元することができることになる。また、HDDはディスク装置の一例に過ぎず、他のディスク装置、例えば、光ディスク装置や光磁気ディスク装置を用いた構成を採用することも可能である。

【0028】

【発明の効果】以上説明したように請求項1に係るディスクアレイ装置によれば、復旧の際に転送が未完となっている場合には、当該格納先アドレス及びサイズに基づき転送が未完となっている領域が検出され、この領域のパリティグループについてディスク装置の記憶内容に基づきパリティデータが作成されて、該当するディスク装置に書き込みが行われる。

【0029】以上説明したように請求項2に係るディスクアレイ装置によれば、複数のディスク装置のいずれかに障害が発生した場合には、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元しても適切な復元がなされる。

【0030】以上説明したように請求項3に係るディスクアレイ装置によれば、複数のディスク装置の特定の1台が、パリティデータを記憶するためのパリティディスクとなっているので、所定のディスク装置にパリティデータを記憶して適切にデータを復元することが可能である。

【0031】以上説明したように請求項4に係るデータ保守方法によれば、複数のディスク装置のいずれかに障害が発生した場合に、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元して適切な復元が可能となる効果がある。

【0032】以上説明したように請求項5に係るデータ保守方法によれば、複数のディスク装置のいずれかにおいて障害が発生したか否かを検出し、障害が発生した場合に、他の正常なディスク装置に記憶されたデータに基づき障害に係るディスク装置に記憶されたデータを復元するので、複数のディスク装置のいずれかにおいて障害が発生した場合にデータの復元がなされることになる。

【図面の簡単な説明】

【図1】本発明の実施の形態に係るディスクアレイ装置の機能ブロック図。

【図2】本発明の実施の形態に係るディスクアレイ装置の詳細ブロック図。

【図3】本発明の実施の形態に係るディスクアレイ装置に用いられるメモリテーブルの内容を示す図。

【図4】本発明の実施の形態に係るディスクアレイ装置

の動作を説明するためのフローチャート。

【図5】本発明の実施の形態に係るディスクアレイ装置の動作を説明するためのフローチャート。

【図6】本発明の実施の形態に係るディスクアレイ装置における電源断から復旧後の動作を説明するための図。

【図7】本発明の実施の形態に係るディスクアレイ装置におけるデータ復元動作を説明するための図。

【図8】従来のディスクアレイ装置の詳細ブロック図。

【図9】RAIDにおけるパリティ生成を説明する図。

【図10】RAIDにおけるデータ書き込み及びパリティ生成を説明する図。

【図11】従来のディスクアレイ装置におけるデータ書き込み時の電源断に際する処理動作を説明するための図。

【図12】従来のディスクアレイ装置におけるデータ復元動作を説明するための図。

【符号の説明】

1 ホスト計算機

2A ディス

クアレイ装置

3 マイクロプロセッサ

4 メモリ

5 プロトコル変換部

6 不揮発性

メモリ

7、8-0～8-4 インタフェース

9 パリティ

回路

10 データバッファ

11-0～1

1-4 HDD

30 ディスク部

31-1～3

1-n ディスク装置

32 データ分割手段

33 パリティ

算出手段

34 転送手段

35 状況記

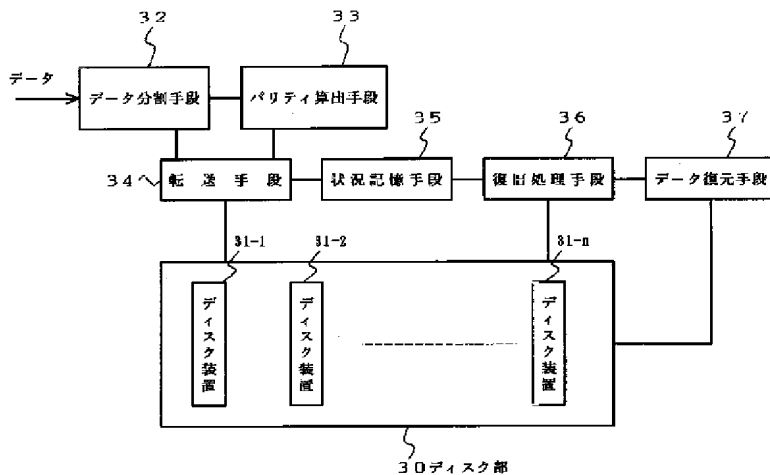
憶手段

36 復旧処理手段

37 データ

復元手段

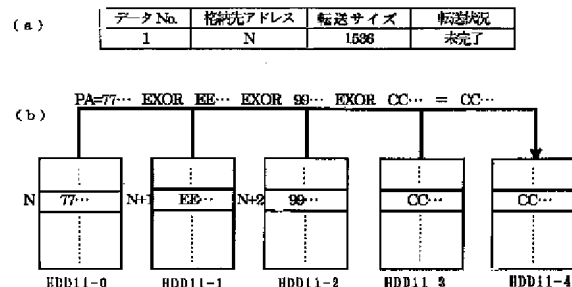
【図1】



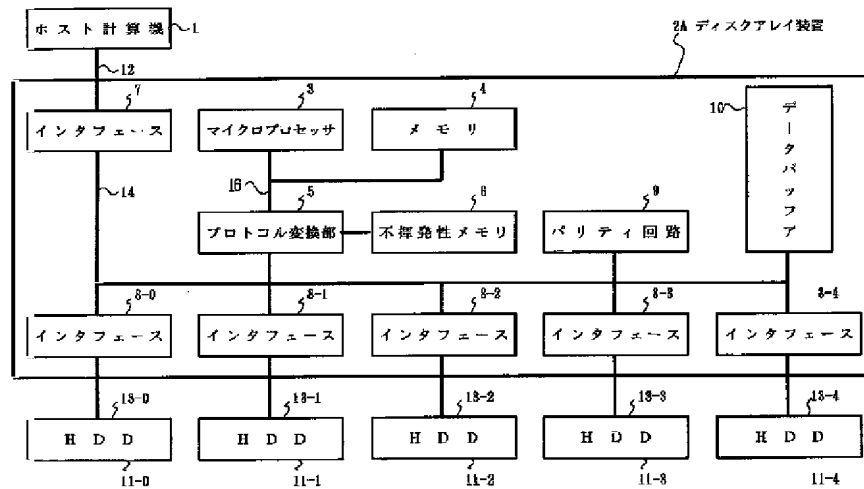
【図3】

書き込み中データ1	格納先アドレス1	転送サイズ1	転送状況1 (完了/未完了)
書き込み中データ2	格納先アドレス2	転送サイズ2	転送状況2 (完了/未完了)
書き込み中データ3	格納先アドレス3	転送サイズ3	転送状況3 (完了/未完了)
書き込み中データ4	格納先アドレス4	転送サイズ4	転送状況4 (完了/未完了)
書き込み中データ5	格納先アドレス5	転送サイズ5	転送状況5 (完了/未完了)
書き込み中データ6	格納先アドレス6	転送サイズ6	転送状況6 (完了/未完了)
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

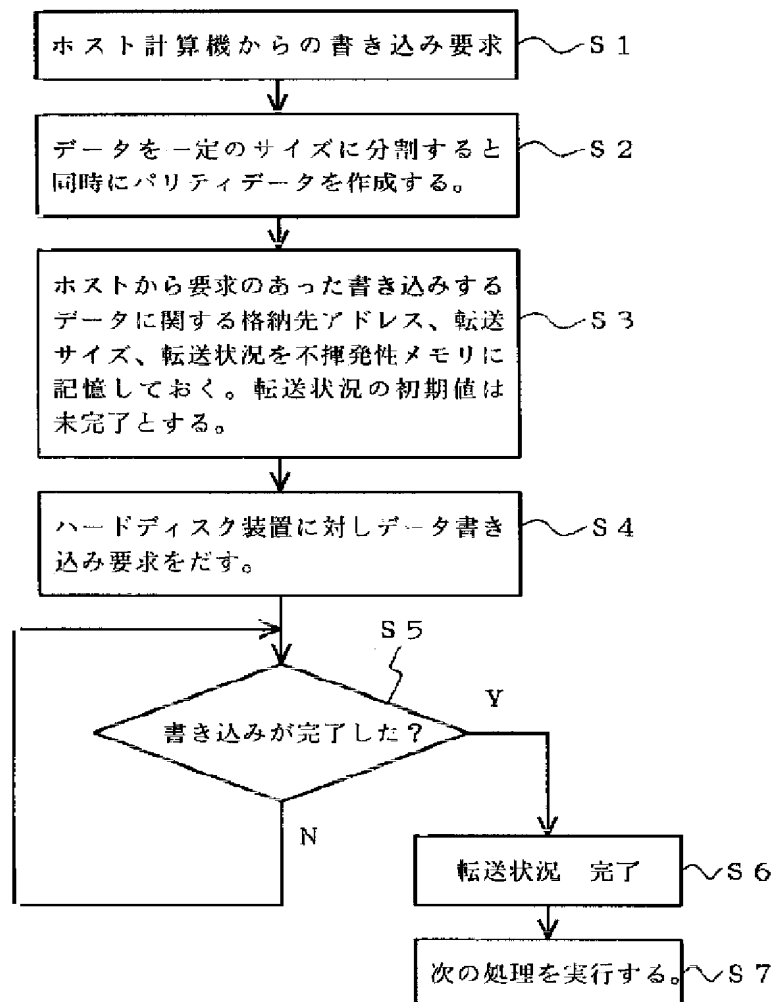
【図6】



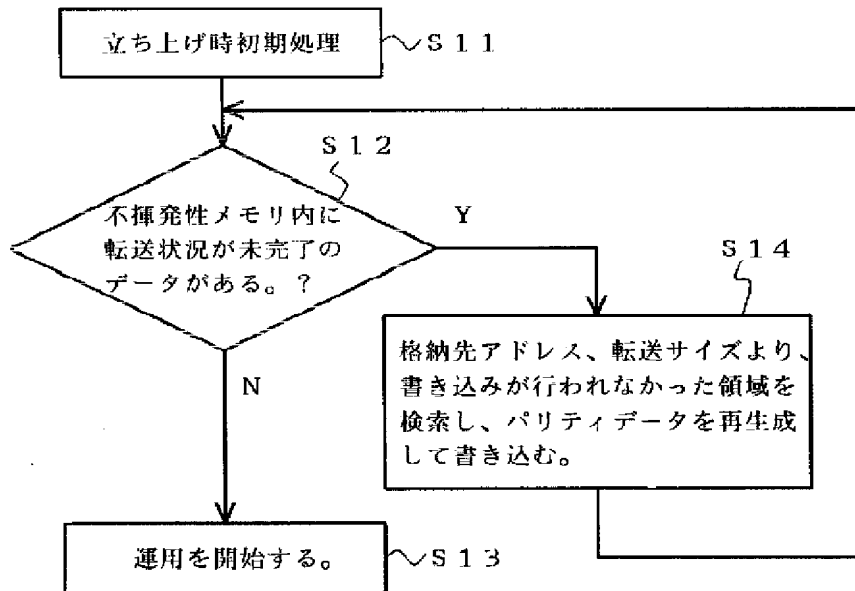
【図2】



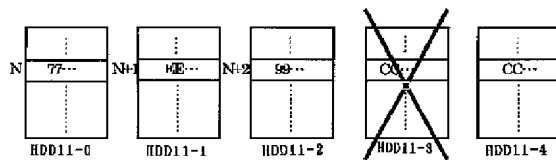
【図4】



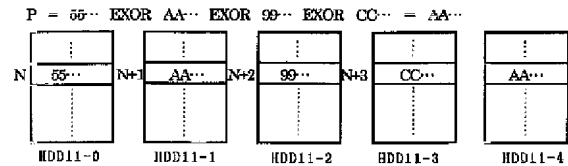
【図5】



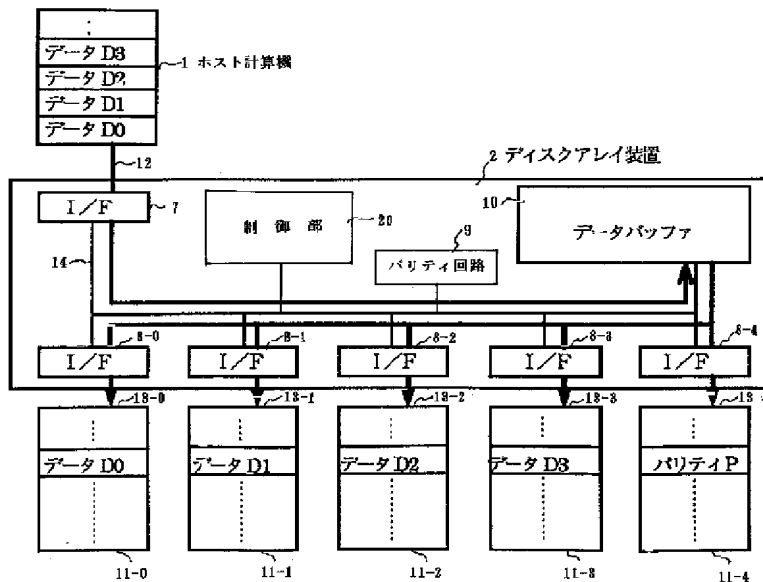
【図7】



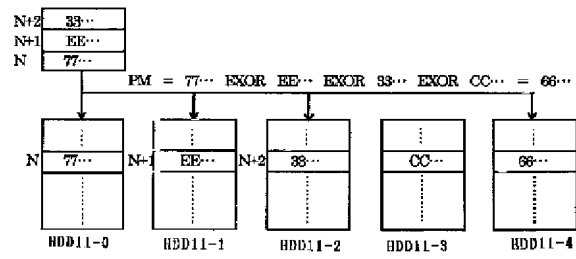
【図9】



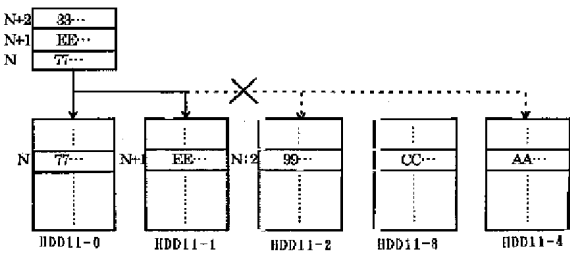
【図8】



【図10】



【図11】



【図12】

